

Comparing 3-D Protein Structures Similarity by Using Fractal Features

Chenyang Cui
State Key Lab. of CAD&CG
Zhejiang University
ccy@cad.zju.edu.cn

Donghui Wang
College of Computer Science and
Technology
Zhejiang University
mii@cs.zju.edu.cn

Jiaoying Shi
State Key Lab. of CAD&CG
Zhejiang University
jyshi@cad.zju.edu.cn

Abstract

In this paper, we propose a new method for finding similarity in 3-D protein structure comparison. Different from the other existing methods, our method is grounded in the theory of fractal geometry. The proposed feature vectors of protein structures are simple to implement. The method is very fast because it requires neither alignment of the chains nor any chain-chain comparison. We calculate the fractal features of a set of 200 protein structures selected from PDB(Protein Data Bank). The experimental result shows that our method is very effective in classification of 3-D protein structures.

1. Introduction

In Protein Databases (e.g. PDB), each year a high number of new 3-D structures of protein molecules is determined either by x-ray crystallography or by NMR techniques or by theoretical prediction. To avoid potential exponential explosion of structures, a basic problem is to which class the new protein molecule belong. Proteins belonging to the same functional family often share common structural features even if there is no evolutionary dependence or sequence similarity between them. So, the protein molecules can be classified by comparing its 3-D structures similarity.

In recent years, many different approaches for extracting protein structural features are developed. Geometric hashing based algorithms [1] and index based algorithm [2] choose a set of reference frames from each target protein and place the other elements of the protein in a hash table, based on each reference frame, the space complexity of this approach is relative to the number of elements considered for each target protein. Tolga Can [3] uses three independent smoothing splines parameterized with respect to the polygonal arc length t to infer the global structure. Similarly, Peter Rogen [4] uses Gauss integrals to analyze and compare protein structure, this method is very fast and used in CATH2.4 database. A popular method is distance matrix [5] and used in FSSP database, in which the

distance is defined as the one between two atoms. Mihael The CE algorithm [6] performs a combinational extension of aligned fragment pairs, it builds an alignment between two protein structures through a combinatorial extension of an alignment path defines by aligned fragment pairs.

The most of above mentioned methods are based on high dimension features. When performing a database search, these methods need exhaustive searching. The challenges outlines in the preceding section motivated us to step away from high dimension features and to instead compare and classify proteins on the basis of their fractal properties.

2. Feature Extraction

2.1. Fractal Background

Proteins are heteropolymers with a variable composition of twenty different amino acids. The amino acid sequence shows that 3-D structure of the protein for the varied composition and nature of their side groups result in a range of possible interactions within the protein. These interactions determine the final structure of protein. So, proteins have an intrinsic self-similarity in the compactness and the packing of their structure. Whether natural or mathematical, all fractals have particular fractal dimensions. These are not the same as the familiar Euclidean dimensions. measured in discrete whole integers such as 1,2,or 3,but a different kind of quantity. Usually non-integer, a fractal dimension indicates the extent to which the fractal object fills the Euclidean dimension in which it is embedded. A protein molecule is made up of one or several polypeptide chains and the protein backbone is a space curve composed of c^α atoms. This motivates us to consider the fractal features of the 3-D protein structures. The following section describes how to extract the fractal features of the protein in our methods.

2.2. Fractal feature extraction

The fractal dimension of a curve may be defined by measuring the length L , with rulers of fixed length ε . The relationship between the length L and length ε is:

$$L(\varepsilon) \propto \varepsilon^{1-D_\varepsilon} \quad (1)$$

here, $D(\varepsilon)$ is the fractal dimension. $L(\varepsilon)$ is defined as the length of the protein molecule chain from the c^α of N polar in polypeptide chain, in which the 1^{st} c^α , $(\varepsilon + 1)^{\text{th}}$ c^α , $(2\varepsilon + 1)^{\text{th}}$ c^α , ..., and $(K\varepsilon + 1)^{\text{th}}$ c^α are connected to a zigzag, c^α is represented as amino acid residue, the number of the amino acid residues is N , if the number of residues n between $(K\varepsilon + 1)^{\text{th}}$ c^α and C polar in polypeptide chain is less than ε , the connected zigzag stopped at the $(K\varepsilon + 1)^{\text{th}}$ c^α . Then, the length of the protein molecule chain $L(\varepsilon)$ is defined as following:

$$L(\varepsilon) = L_z(\varepsilon) + \frac{n}{\varepsilon} \cdot \frac{L_z(\varepsilon)}{K} \quad (2)$$

where, $1 \leq \varepsilon \leq N - 1$, and $L_z(\varepsilon)$ is the length of the connected zigzag, and $\frac{n}{\varepsilon} \cdot \frac{L_z(\varepsilon)}{K}$ is the length of the remainder whose c^α . By the curve approximation, we can compute the slope of the fractal curve ($\text{Log}(L(\varepsilon)) - \text{Log}(L(\varepsilon))$) $S(\varepsilon)$, then,

$$D(\varepsilon) = 1 - S(\varepsilon) \quad (3)$$

So, the feature of the protein structure can be defined as $D(\varepsilon)$.

3. Experiment results and discussion

In our experiments, we downloaded about two hundred protein structures from PDB. Experiment results show that the fractal dimension feature performs well in comparing 3-D protein structure similarity. See Figure 1, the backbones of several protein molecules are shown and the corresponding fractal dimensions are given.

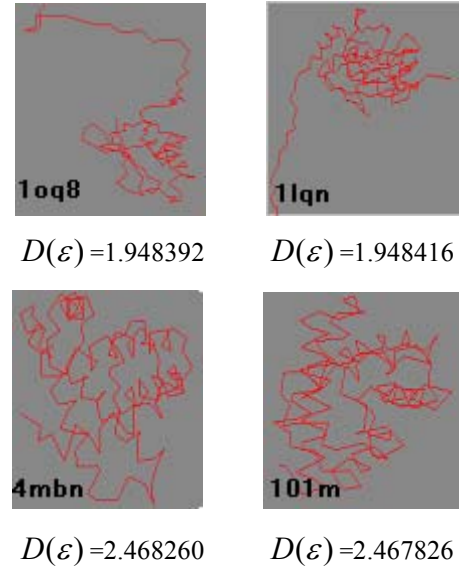


Figure 1. The backbone of the protein molecule

From the experiments, we can find the differences between the fractal dimension of the two pairs protein molecules are very small. So, the fractal dimension can be considered as the feature of the 3-D protein structure. Obviously, $D(\varepsilon)$ is sensitive to the fixed length ε , if ε is in certain region, the curve is linear. In the future, we will improve the robust of $D(\varepsilon)$ to avoid its sensitivity to ε .

References

- [1] [1] S.C. Chen, T.H. Chen, "Protein Retrieval By Matching 3D surfaces.", GENSIPS 2002, Raleigh, North Carolina, USA., October 2002.
- [2] Z. Aung, W. Fu, K.Lee, and Tan., "An Efficient Index-based Protein Structure Database Searching Method". In Proc. Of the 8th International Conference on Database Systems for Advanced Application(DASFAA),2003.
- [3] T. Can, Y.F. Wang, "CTSS: A Robust and Efficient Method for Protein Structure Alignment Based on Local Geometrical and Biological Features", Proceedings of the computational Systems Bioinformatics (CSB'03).
- [4] P. Rogen, B. Fain, "Automatic classification of protein structure by using Gauss Integrals", PNAS, Jan.7,2003, vol.100,No.1,119-124.
- [5] J. Holm, C. Sander, "Protein Structure Comparison by Alignment of Distance Matrices", Journal of Molecular Biology,233(1):123-138,1993
- [6] H.N.Shindyalov, P.E. Bourne, "Protein structure alignment by incremental combinational extension(CE) of the optimal path", Protein Engineering,11(9):739-747,1998